White Paper Report

Report ID: 98609

Application Number: PK5002207

Project Director: Gregory Crane (gregory.crane@tufts.edu)

Institution: Tufts University

Reporting Period: 10/1/2007-8/31/2011

Report Due: 11/30/2011

Date Submitted: 12/5/2011

Scalable Named Entity Identification in Classical Studies
NEH Award PK-50022-07
Final White Paper

November, 2011
Gregory R. Crane, Tufts University

Introduction

The names of people, places, ethnic groups, months of the year and other named entities are critical components for many, if not most, textual sources. A great deal of work has gone into providing services that can identify named entities in English and other modern languages. While much has been accomplished for Greek and Latin by working with English translations of original source texts, no services have been available to identify names in the Greek and Latin sources themselves.[1]

A large amount of work has gone into the development of named entity services in general. Wikipedia now contains cross-references between articles in different languages, providing, in effect, a growing multilingual thesaurus for personal names across a number of languages[2] – the entry for Thucydides[3](the name of a major historian and an important Athenian statesman among other people) includes versions of the name in 70 languages; the article for Plato[4] includes versions of the Greek name (which describes a comic poet as well as the famous philosopher) in 160 languages. In addition, the Wikipedia derivative, DBpedia[5] provides this data in a format designed for reuse.

This paper focuses upon how we ultimately addressed the problem of providing the language specific services needed to identify names in the original Greek and Latin sources. Our goal is not to provide a full suite of named entity identification services but is instead to provide the language and domain specific services needed such that Greek and Latin can benefit from more general services.

Our work ultimately focused on three tasks: (1) annotating names in Greek and Latin sources directly (as opposed to using English-language services to annotate English translations of Greek and Latin sources); (2) publishing machine actionable versions of two historical encyclopedias, Smith's 1854 *Dictionary of Greek and Roman Geography* and Smith's 1873 *Dictionary of Greek and Roman Biography and*

[1] While the importance of both named entities and automatic named entity identification within historical texts has grown enormously over the last five years, little work has been done in terms of historical languages such as Greek or Latin. For some recent interesting named entity identification work with English language cultural heritage materials, see Grover et al. (2010) in terms of place names, and for personal names see Clough et al. (2009).
[2] In fact, the Wikipedia corpus is frequently used as a data source for testing named entity identification and disambiguation algorithms across languages. For some recent work, see Balasuriya et al. (2009) and Waltinger and Mehler (2008)
[3] http://en.wikipedia.org/wiki/Thucydides
[4] http://en.wikipedia.org/wiki/Plato
[5] http://dbpedia.org

Mythology with as many citations as possible automatically identified; (3) creating a database in which more than 100,000 entries extracted from print back-of-the-book indices have been converted into machine actionable form. This paper describes each of these three activities and then concludes by describing materials that will be published on the Perseus Web Site[6] in 2012.

1. Annotated Greek and Latin Corpora

We have been able to add named entity markup to the open source corpora of Greek and Latin distributed by the Perseus Digital Library. Table 1 summarizes current coverage of names.

	Tokens	Names	MorphAnalysis		NameClass	
Greek	11,136,261	578,451	544,180	94.1%	507,231.00	87.7%
Latin	6,437,607	351,567	323,582	92.0%	300,746.00	85.5%

Table 1: Coverage of Greek and Latin names analyzed and classified

The open-source Greek and Latin corpora currently contain 11.1 and 6.4 million words of text respectively. The percentage of proper names is comparable in each corpus (5.1% vs. 5.4% of the words in the Greek and Latin corpora respectively).

We were able to generate morphological analyses for 94% of the Greek and 92% of the Latin names. In addition, we have been able to generate semantic classifications (e.g., person vs. place vs. ethnic group) for 87.7% and 85.5% of the Greek and Latin names in the corpus as a whole. (Note that we at present only classify those names for which we have morphological analyses).

[6] http://www.perseus.tufts.edu

```
e)n <placeName key="0001314:*xerso/nhsos:place:fem:dat">*xerronh/sw|</placeName>.
e)pe/skhye de/, e)a/n ti pa/qh|, ta/lanton <milestone unit="reiskpage" n="504"/>
<pb id="p.37"/>
me\n e)pidou=nai th=| gunaiki\ kai\ ta\ e)n tw=| dwmati/w| dou=nai,
ta/lanton de\ th=| qugatri/. kate/lipe <add>de\</add> kai\ ei)/kosi
mna=s th=| gunaiki\ kai\ tria/konta stath=ras <placeName
key="0018237:*ku/zikos:place:masc:acc 0017237:*kuzikhno/s:ethnic:masc:acc
0001391:*ku/zikos:group:masc:acc">*kuzikhnou/s</placeName>.
tau=ta de\ pra/cas kai\ oi)/koi a)nti/grafa katalipw\n w)/|xeto
strateuso/menos meta\ <persName
key="0000660:*qra/sullos:person:masc:gen">*qrasu/llou</persName>. a)poqano/ntos de
e)kei/nou e)n <placeName key="0043281:*)/efesos:place:masc:dat
0043281:*)/efesos:place:fem:dat 0000453:*)/efesos:person:masc:dat
0000453:*)/efesos:person:fem:dat">*)efe/sw|</placeName> <persName
key="0001000:*diogei/twn:person:masc:nom/voc">*diogei/twn</persName> th\n me\n
qugate/ra e)/krupte
to\n qa/naton tou= a)ndro\s kai\ ta\ gra/mmata lamba/nei,
a(\ kate/lipe seshmasme/na, fa/skwn ta\ nautika\ xrh/mata
dei=n e)k tou/twn tw=n <milestone unit="reiskpage" n="505"/>grammatei/wn
komi/sasqai. e)peidh\
de\ xro/nw| e)dh/lwse to\n qa/naton au)toi=s kai\ e)poi/hsan
ta\ nomizo/mena, to\n me\n prw=ton e)niauto\n e)n <placeName
key="0000209:*peiraieu/s:place:masc:dat">*peiraiei=</placeName>
dih|tw=nto: a(/panta ga\r au)tou= katele/leipto ta\ e)pith/deia.
e)kei/nwn d' e)pileipo/ntwn tou\s me\n pai=das ei)s a)/stu
a)nape/mpei, th\n de\ mhte/ra au)tw=n e)kdi/dwsin e)pidou\s
pentakisxili/as draxma/s, xili/ais e)/latton w(=n o( a)nh\r
au)th=s e)/dwken. o)gdo/w| d' e)/tei dokimasqe/ntos meta\ tau=ta
tou= presbute/rou toi=n meiraki/oin kale/sas au)tou\s ei(=pe
<persName key="0001000:*diogei/twn:person:masc:nom/voc">*diogei/twn</persName>,
```

Figure 1: Sample named entity tagging for Greek (from Dionysius of Halicarnassus). The key attributes contain the size of the corresponding entries in the appropriate Smith encyclopedias. The lengths of all articles for "Alexandria," for example, in the Smith biographical encyclopedia, for example, are added and then compared to the aggregate lengths of all articles for Alexandrias in the Smith geographic encyclopedia. If the biography entries are longer, our default assumption is that we have a personal name, but if the geographic articles are longer, we assume by default that it is a place name.

```
promissisque accendens, cur resumpsissent arma, <name type="noclass"
key="0000001:Pannonicus:noclass:fem:acc">Pannonicas</name> legiones interrogabat:
illos esse campos, in quibus
abolere labem prioris ignominiae, ubi reciperare gloriam
possent. tum ad <name type="noclass"
key="0000001:Moesicus:noclass:masc:acc">Moesicos</name> conversus principes
auctoresque
belli ciebat: frustra minis et verbis provocatos <name type="group"
key="0000553:Vitellianus:group:masc:acc
0000001:Vitelliani:noclass:masc:acc">Vitellianos</name>,
si manus eorum oculosque non tolerent. haec, ut quosque
accesserat; plura ad tertianos, veterum recentiumque admo- <pb ed="oct"/>
nens, ut sub <persName><foreName n="Marcus">M.</foreName> <surname
n="0125991:Antonius:person:neut:dat 0125991:Antonius:person:neut:abl
0125991:Antonius:person:masc:dat
0125991:Antonius:person:masc:abl">Antonio</surname></persName> <name type="ethnic"
key="0006686:Parthus:ethnic:masc:acc  0006686:Parthus:ethnic:fem:acc
0000001:Parthi:noclass:masc:acc">Parthos</name>, sub <persName><surname
n="0005069:Corbulo:person:masc:abl">Corbulone</surname></persName> <name
type="ethnic" key="0026658:Armenius:ethnic:masc:acc
0002083:Armenius:group:masc:acc">Armenios</name>,
nuper <name type="ethnic" key="0001000:Sarmata:ethnic:masc:acc
0000001:Sarmatae:noclass:masc:acc">Sarmatas</name> pepulissent. mox infensus
praetorianis 'vos'
inquit, 'nisi vincitis, pagani, quis alius imperator, quae castra alia excipient?
illic signa armaque vestra sunt, et mors
victis; nam ignominiam consumpsistis.' undique clamor, et
orientem solem (ita in <placeName key="0084099:Syria:place:fem:nom/voc
0084099:Syria:place:fem:abl 0000866:Syria:person:fem:nom/voc
0000866:Syria:person:fem:abl  0000001:Syrius:noclass:neut:nom/voc/acc
```

Figure 2: an example of a Latin text with named entity annotation.

1.1 Identifying Proper Names

Greek and Latin editions use capitalization to mark proper nouns and we drew upon this in our annotation workflow. In preparing the corpora for named entity analysis, we needed to distinguish proper names from common nouns that had been capitalized for various reasons. Conventions of capitalization vary across different editions of Greek and Latin. Editors variously capitalize words at the start of different text locations including paragraphs, individual sentences, quotations, metrical lines and in other varying circumstances. In order to identify names, we needed to distinguish where these capitalized words were and were not actually names.

To do so we drew upon the corpora as a whole. We identified upper case words that appeared directly after another word without intervening punctuation (i.e., words in the middle of a sentence). We then counted the number of times that a word appeared in such mid-sentence contexts both in upper and in lower case (e.g., how often *Commodus* vs. *commodus* appeared in Latin). If the word appeared more often in upper case in the middle of a sentence, we assumed that it was a proper noun whereas if it appeared more often in lower case in the middle of the sentence, we assumed that it was a common noun.

1.2 Morphological Analysis

In languages such as English and Chinese we can annotate names in sample texts and thus train classifiers to recognize names in unseen texts. The highly inflected nature of Greek and Latin, however, makes relatively straightforward approaches utilized for English language texts ineffective: we need to know not only that Caesar is a personal name but that all the inflected forms of this word (e.g., *Caesaris*, *Caesare*, *Caesari*, *Caesarem*) are forms of the same word. Thus, we need not only to recognize that *Caesare* is a personal name but also that it is a form of *Caesar*.

Morphological analysis is a fundamental task that advances any named entity identification system. By including the morphological analyses for individual names, we provide third parties with better data with which to apply their own systems for subsequent named entity classification.

Digital Classicists have addressed the problem of Greek and Latin morphology since David Packard began early work in the 1970s (Packard 1973). We built upon Morpheus, a morphological analyzer for Greek and Latin originally developed in the late 1980s by Gregory Crane (Crane 1991), who in turn built upon code from Neel Smith and Joshua Kosman (then graduate students at Berkeley). This morphological analysis system, complicated as it may be, is relatively straightforward in design and the same C Programming Language code base has been effective for more than 20 years, running without modification on a range of systems.

The great challenge with Greek and Latin morphology is the lexicon.[7] To analyze highly inflected languages such as Greek and Latin we need to manage databases of stems that tell the analyzer what forms are legal for a particular form. In effect, the morphological complexity that has tormented school children for thousands of years makes the challenge of managing a database of stems challenging. Heap's Law[8] models the degree to which the vocabulary of a language gradually and inexorably grows over time. The lexicon for a living language is never complete. Greek and Latin may no longer be living languages but they remained a medium of communication for thousands of years and even the classical periods of Greek and Latin extend approximately 800 and 1400 years respectively (750 BCE through 650 CE, 200 BCE through 600 CE respectively). Names in particular change more rapidly than the core terms of Greek and Latin – while grammarians tried to systematize these languages no one could prevent the names of new people, places, ethnic groups and organizations from entering into them.

We mined digitized versions of Greek and Latin lexica for the stems upon which morphological analysis depends. This was no small task – and, in fact, our justification for entering the first Greek lexicon more than two decades ago (the 19th century *Intermediate Greek-English Lexicon* by Liddell and Scott[9]) was the necessity

[7] For more on the importance of computational lexicons and morphology in Greek and Latin, see Bamman and Crane (2009).
[8] http://en.wikipedia.org/wiki/Heaps%27_law
[9] http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3atext%3a1999.04.0058

of building up a database of stems. Even when we had a transcribed version of the print text (an expensive task in and of itself), identifying and analyzing the morphological data was no small task – the morphology often appears in forms that are abbreviated, ambiguous, inconsistent, and complex. The more irregular the form, the more complicated the entry and thus the more difficulties that arose as we tried to sift machine actionable data from the tangled words and abbreviations of print.

The main Greek lexicon, the ninth edition *Liddell Scott Jones* (LSJ)[10], only covers a handful of proper nouns – just 2,600 of 116,000 entries cover proper nouns. Coverage in the Lewis and Short *Latin Dictionary*[11] is much better (8,500 of 51,500 entries describe proper names) but this provides only an initial starting point. Neither of these lexica, however, provides an easy mechanism whereby we can determine whether a name is a person, place or some other category. In order to provide better coverage for Greek and Latin names, we thus extracted names from the Smith's biographical and geographical dictionaries (described below).

As new texts from different domains are added to a corpus, new names inevitably appear. A small number of high frequency names for which no morphological analysis is available will appear as new texts are added. Thus, the addition of later Greek historians introduced names such as the Greek forms of Constantius and Valentinianus for which no morphological profile existed. High frequency names without morphological analyses can be rapidly identified and added to the morphological database. A few minutes is usually enough to update the database to cover the most common new names that appear in a new source text.

In practice, however, most of the names for which no morphological analysis exists appear very infrequently. The 34,371 Greek and 27,985 Latin names in running text without morphological analyses represent 19,094 and 12,995 forms.

1.3 Classifying Names

	Person	Place	Ethnic	Other	Unclassified
Greek	272,422	103,656	115,449	49,975	36,949
Latin	201,297	54,711	33,215	39,508	22,836

Table 2: Major Categories of Names

The nature of Greek and Latin make semantic classification of names easier than in English. When we compare the headwords in the Smith geographical and biographical dictionaries, more than 97% of the names appear in only one list or the other. Where English names such as Washington and London can apply to people or places, Greek and Latin often use suffixes to distinguish people from the places named after them (e.g., Alexander and Alexandria, Constantine and Constantinople).

[10] http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3atext%3a1999.04.0057
[11] http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3atext%3a1999.04.0059

Some names are inherently ambiguous (the name Oceanus can describe either the Atlantic ocean or the sea god Oceanus) but for the most part semantic classification is relatively straightforward.

Where names are ambiguous, we model the relative frequency of the two classes, using the length of the entry in the biographical versus the geographical dictionaries as an initial value: if the biographical entry for a word is longer than the geographic entry, we assume that the name designates a person, but we encode the relative size of the two entries in the annotation so that we can later review ambiguous decisions.

This relatively simple approach yields classification accuracies that approach 90% and is comparable to the results of more sophisticated systems. Our goal is to provide preliminary classifications with the expectation that others will apply their own methods. Within the context of named entity classification, the annotated Greek and Latin corpus provides a demonstration collection but the Smith encyclopedias and the indices (discussed below) provide the data that will most benefit subsequent work.

Additionally, the category *other* includes a small number of classes, including buildings, constellations, festivals and titles. Most of the names that we have classified are, however, for people, places, and ethnic groups. People are the most prominent category but the greater gap between people and places observed in Latin stands out. Our preliminary hypothesis is that the Greek collection now contains a greater percentage of historical works with a greater emphasis on places but this phenomenon deserves further study.

Greek (Anglicized)		**Latin**	
8,878	Zeus	9,217	Dominus
4,252	Alexander	3,917	Caesar
4,178	Caesar	6,070	Deus
3,157	Socrates	2,540	Iuppiter
2,874	Philip	2,362	Cicero
2,576	Heracles	1,673	Pompeius
2,321	Apollo	1,471	Christus
2,229	Homer	1,277	Antonius
2,075	Dionysos	1,246	Claudius
1,923	Pompey	1,232	Africanus

Table 3: The most common personal names in the Greek and Latin collections. (Greek names have been Anglicized.)

Greek (Anglicized)		**Latin**	
2804	Rome	3362	Roma
2677	Greece	2676	Israel
2473	Athens	1857	Italia

2310	Egypt	1007	Aegyptus
2298	Carthage	972	Gallia
2073	Italy	941	Asia
1926	Asia	916	Hierusalem
1770	Sicily	902	Sicilia
1302	Libya	758	Graecia
1125	Peloponnesus	664	Hispania

Table 4: The most common placenames in the Greek and Latin collections. (Greek names have been Anglicized.)

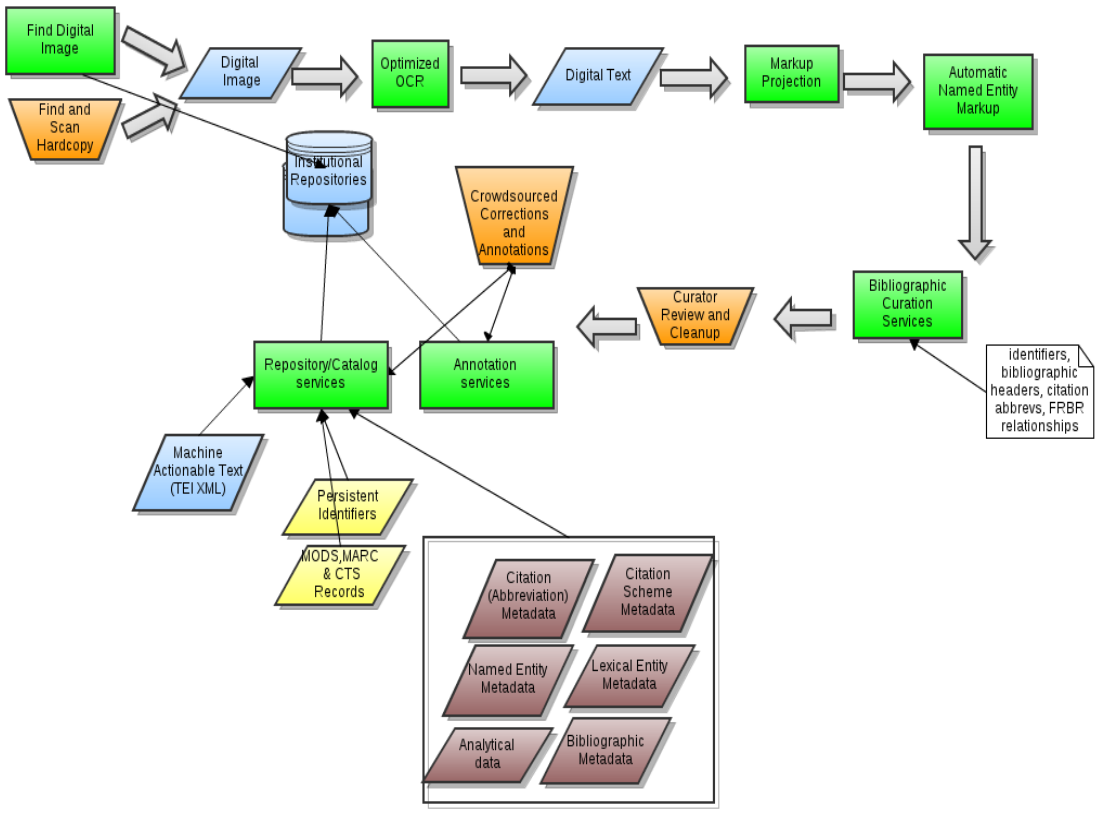

Figure 3: Target Annotation Workflow

Finally, Figure 3 above describes a scalable workflow that takes advantage of automated processing to seed annotations of named entities and enables crowd-sourcing of the manual curation of the automatically produced annotations and classifications.

2. The Digital Smith Encyclopedias of People, Places and Daily Life.

As discussed in the previous section, the Smith's 1854 *Dictionary of Greek and Roman Geography* and 1873 *Dictionary of Greek and Roman Biography and Mythology*, which contain 10,000 and 20,000 entries respectively, have been utilized in this project to assist in the automatic classification of place and personal names.

The scope of each is comparable, as indicated by their prefaces:

"Separate articles are given to the geographical names which occur in the chief classical authors, as well as to those which are found in the Geographers and Itineraries, wherever the latter are of importance in consequence of their connection with more celebrated names, or of their representing modern towns,--or from other causes. But it has been considered worse than useless to load the work with a barren list of names, many of them corrupt, and of which absolutely nothing is known." (Smith, *Geography*, p. viii).

The biographical encyclopedia promises greater coverage, including Byzantine as well as classical authors:

"The biographical articles in this work include the names of all persons of any importance which occur in the Greek and Roman writers, from the earliest times down to the extinction of the Western Empire in the year 476 of our era, and to the extinction of the Eastern Empire by the capture of Constantinople by the Turks in the year 1453." (Smith, *Biography*, p. vii).

In addition, we are distributing the XML source for the *Dictionary of Greek and Roman Antiquities* (1890).

```
<div2 type="entry" id="thucydides-bio-2"><head><persName
lang="la"><surname>Thucy'dides</surname></persName></head>

<p>2. A Pharsalian, was a proxenus of the Athenians and happened to be at Athens in
<date value="-411">B. C. 411</date>, during the usurpation of the Four Hundred.
When the tumult against the government broke out in the Peiraeeus, and Theramenes
had gone thither with promise of quelling it, Thucydides with some difficulty
restrained the adherents of the oligarchs <pb n="1112"/> in the city from marching
down to attack the rioters, representing the mischief attendant on civil discord
while the Lacedaemonians were so close at hand. (<bibl n="Thuc. 8.92">Thuc.
8.92</bibl>.)</p></div2>


<div2 type="entry" id="thucydides-bio-3"><head><persName
lang="la"><surname>Thucy'dides</surname></persName></head>

<p>3. A lieutenant of Martius Verus, by whom he was sant to establish Soaemus on
the throne of Armenia, in the reign of M. Aurelius Antoninus. Thucydides
accomplished his mission. (Suid. s. <hi rend="ital">s. v.</hi> <foreign
lang="greek">*Ma/rtios</foreign>; see above, Vol. I. p. 363a.) </p><byline>[<ref
target="author.E.E">E.E</ref>]</byline></div2>


<div2 type="entry" id="thucydides-bio-4"><head><persName lang="la"
id="tlg-0003"><surname>Thucy'dides</surname></persName></head>

<p>(<label lang="greek">*Qoukudi/dhs</label>), the historian, belonged to the demos
Halimus, and Halimus belonged to the Leontis.
He simply calls himself an Athenian (<bibl n="Thuc. 1.1">Thuc. 1.1</bibl>). His
father's name was Olorus (4.104). Marcellinus, and some other later writers, say
that the name was Orolus.
The two forms are easily confounded, and we assume the true name to be Olorus.
Herodotus (<bibl n="Hdt. 6.39">6.39</bibl>) mentions a Thracian king called Olorus,
whose daughter Hegesipyle married Miltiades, the conqueror of Marathon, by whom she
became the mother of Cimon.
```

Figure 4: Basic markup for the Smith encyclopedia. Many – but by no means all – primary source citations are automatically marked. Ambiguous citations (e.g., "4.101" rather than "Thuc. 4.101") are not annotated. Works with inconsistently formatted citations are often not annotated (e.g., Suidas). The figure above shows dates, page breaks, paragraphs, articles, article id-s, links to authority lists such as the Thesaurus Linguae Graecae (TLG) canon where available, names of article authors etc.

In order to extract machine actionable data from the digital versions of these works, we needed to perform a number of tasks:

1) Identification of article boundaries and headings: Typographical cues such as capitalization provide a starting point but the formats of article names vary, especially when describing people with names of various complexity. We largely based headers on the first major name and then used numbers to distinguish, for example, the many different places named Alexandria or

people named Antonius. In addition, not all entries map to single entities. Rivers, mountains, towns, and other features may have their own entries or be part of larger entries. We also find biographical articles that describe several figures (especially where an article describes several figures of the same name that cannot precisely be distinguished).

2) Identification of citations to primary sources: So far more than 33,000 and 43,000 print citations to Classical authors have been identified in the geographical and biographical dictionaries. Many more, especially to later Classical authors, remain to be identified. On the average, each person or place mentioned in these dictionaries has between two and three machine actionable links to primary sources. This data provides a starting point for services that can automatically associate names with particular entities (e.g., Alexander with Alexander the Great, Alexander of Aphrodisias, etc.).
3) Additional markup: We also provide versions of the Smith Encyclopedias that have more extensive automatic markup, annotating dates, possible family relationships (e.g., "son of"), occupations (e.g., "writer," "sculptor") and other features to support more advanced information extraction. These files also include specialized English named entity tagging that considers the relative frequency of articles within the Smith Encyclopedia series.

```
P.S</ref>]</byline></div2>
 <div2 type="entry" id="pericleitus-bio-2"><head><persName n="0.9:00000182:pericleitus-bio-2@
0.08:00000734:pericleitus-bio-1"><surname>Pericleitus</surname></persName></head>
 <p>artist. [PERICLYTUS.]</p></div2>
 <div2 type="entry" id="pericles-bio-1"><head><persName n="0.99:00061510:pericles-bio-1@0.00:
00000456:pericles-bio-2"><surname>Pericles</surname></persName></head>
 <p>(<label lang="greek">*Periklh=s</label>).</p>
 <p>1. The greatest of Athenian statesmen, was the <rs type="family-relations" n="son-of">son
 of</rs> <persName n="0.54:00001823:xanthippus-bio-4@0.11:00000391:xanthippus-bio-6@0.10:000
00343:xanthippus-bio-1@0.09:00000312:xanthippus-bio-5@0.08:00000282:xanthippus-bio-3@0.06:00
000202:xanthippus-bio-2">Xanthippus</persName>, under whose command the victory of Mycale wa
s gained, and of Agariste, the great <rs type="family-relations" n="grand-daughter-of">grand
-daughter of</rs> <persName n="0.48:00004261:cleisthenes-bio-2@0.40:00003557:cleisthenes-bio
-1@0.11:00001048:cleisthenes-bio-3">Cleisthenes</persName>, <rs type="organizational relatio
ns" n="tyrant-of">tyrant of</rs> <placeName n="0.98:00032277:sicyon-geo@0.01:00000587:sicyon
-bio-1">Sicyon</placeName>, and niece of Cleisthenes, the <rs type="organizational relations
" n="founder-of">founder of</rs> the later Athenian constitution. (<bibl n="Hdt. 6.131">Hdt.
 6.131</bibl> ; <abbr expan="confer">comp.</abbr> CLEISTHENES.) Both Herodotus (<abbr expan=
"loco citato">l.c.</abbr>) and Plutarch have thought the story, that before <rs type="family
-relations" n="is-birth">his birth</rs> <rs type="family-relations" n="is-mother">his mother
</rs> dreamed that she gave birth to a lion, of sufficient interest to deserve recording. Pe
ricles belonged to the deme Cholargos in the tribe Acamantis. The date of <rs type="family-r
elations" n="is-birth">his birth</rs> is not known. The early period of his life was spent i
n retirement, in the prosecution of a course of study in which his noble genius found the mo
```

Figure 5: More advanced markup. This includes named entity identification, a range of social relationships, events, and occupations. The added markup makes the file cumbersome for some systems and this version is thus distributed separately.

The fact that we had two distinct reference works was particularly helpful. Unlike English (especially American English), Greek and Latin do not normally use the same form to describe people and places (e.g., Washington or Lowell as names of

people or places). When we compared entries in the geographical and biographical encyclopedias, we found that more than 97% of the entries occurred in only one or the other and that there was relatively little semantic ambiguity. These two works provide a powerful initial starting point for downstream services that analyze the vocabulary and the citations within individual articles to find additional instances of these names.

3. Machine actionable indices (> 100,000 entries)

The Smith biographical and geographical encyclopedias provide broad coverage for people and places that appear in surviving Greek and Latin sources but they do not set out to provide detailed coverage for major figures or to cover many less important figures that appear in individual sources. To provide this greater detail, we have created a testbed from digitized back-of-the-book indices.

```
<div1 type="entry" xml:id="abioi-cunliff"><head><foreign xml:lang="greek">*)/abioi</foreign>
</head>
<p><foreign xml:lang="greek">(dikaio/tatoi a)nqrw/pwn).</foreign> An unknown race so describ
ed <bibl n="Hom. Il. 13.6">Il. 13.6</bibl>.
</p></div1>

<div1 type="entry" xml:id="ablerus-cunliff"><head><foreign xml:lang="greek">*)/ablhros</fore
ign></head>
<p>T. S. <ref target="#antilochus-cunliff">Antilochus</ref> <bibl n="Hom. Il. 7.32">Il. 7.32
</bibl>.
</p></div1>

<div1 type="entry" xml:id="abydus-cunliff"><head><foreign xml:lang="greek">*)/abudos</foreig
n></head>
<p>A city on the s.e. of the Hellespont, a little below <ref target="#sestus-cunliff">Sestus
</ref> on the other side <bibl n="Hom. Il. 2.836">Il. 2.836</bibl>.- <foreign xml:lang="gree
k">*)abudo/qen [-qen],</foreign> from A. <bibl n="Hom. Il. 4.500">Il. 4.500</bibl>.-<foreign
 xml:lang="greek">*)abudo/qi [-qi],</foreign> at A. <bibl n="Hom. Il. 16.584">Il. 16.584</bi
bl>.
</p></div1>

<div1 type="entry" xml:id="agathon-cunliff"><head><foreign xml:lang="greek">*)aga/qwn</forei
gn></head>
<p>(<foreign xml:lang="greek">di=os</foreign>). A son of Priam (his mother not named) <bibl
n="Hom. Il. 24.249">Il. 24.249</bibl>.
</p></div1>
```

Figure 6: Part of the machine actionable version of the Cunliffe Index of people and places for the Homeric epics. The markup associates particular entities with a name and with passages within the Homeric epics.

The indices essentially map three different features:

1) For each entry, they provide a unique identifier (e.g., “abydus-cunliffe” above). These identifiers are unique and describe entities within the context of a particular work.

2) For each entry, they provide one more Greek or Latin names, which provide information whereby we can identify instances of those entities in actual text (e.g., “Caesar” can be matched against “Caesarem,” “Caesaris,” etc.)
3) For each entry, they provide (where these are available) machine actionable citations (e.g., “Il. 2.836” above). These allow us to disambiguate references to different people and places with the same name.

Each back-of-the-book index describes a unique and disjoint space – the index for Herodotus is separate from that for Thucydides that is, in turn, separate from that for Homer. The people and places are, however, *not* disjoint – the same people and places can appear in multiple works and in multiple entries. To create a unified index, we need to determine when, for example, alexander-12 in one index corresponds to alexander-6 in another.

There are at least two methods to address this challenge.

1) We can analyze the vocabulary of each passage where one entity (e.g., Alexander the Great) occurs and then compare it with the vocabulary of a different entry (e.g., Alexander of Aphrodisias). We can then use the vocabulary profiles to determine whether subsequent references to Alexander more likely refer to Alexander the Great, Alexander of Aphrodisias or some other Alexander.
2) We can look for instances where an entry in the Smith biographical or geographical encyclopedia cites the same passage as a back-of-the-book index. If the names match (e.g., each is an Alexander), we can infer that the Smith entry and the book entry describe the same entity. We can then unify the citations in the Smith article with those in the author article, linking, in effect, two sub-networks into a single network about a single author.

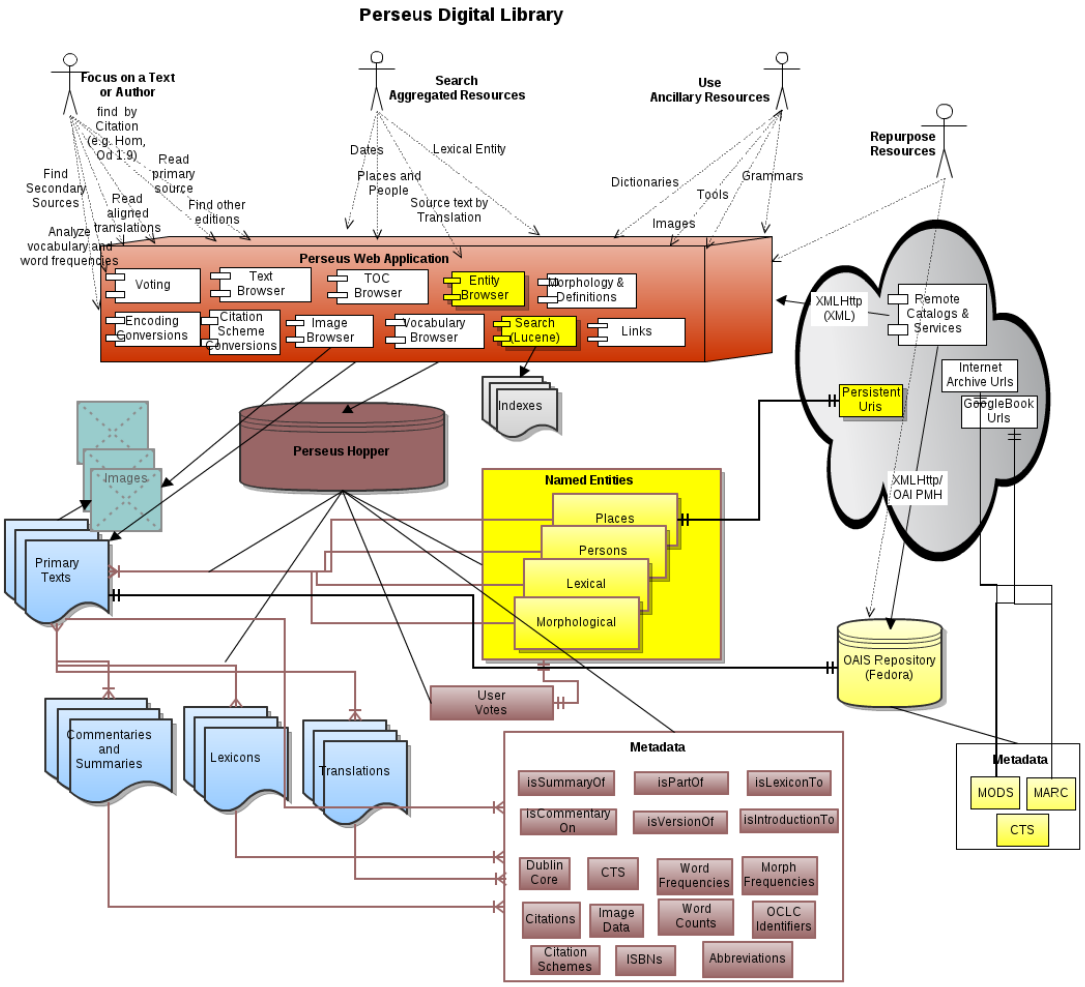

Figure 7: Perseus Digital Library

Figure 7 describes at a high-level the data and functionality provided by the Perseus Digital Library. Named Entity related components are highlighted in yellow.

Conclusion

Support from the NEH/IMLS program allowed us to develop core elements of infrastructure necessary for working with named entities in Greek and Latin sources – a foundation capacity for any cyber-infrastructure. We will be able to publish these datasets, converted into TEI P5 and Unicode with support from the DFG/NEH Hellespont project, under a Creative Commons license on the Perseus site early in 2012.

References

Balasuriya, Dominic, Nicky Ringland, Joel Nothman, Tara Murphy, and James R. Curran. "Named Entity Recognition in Wikipedia." *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, 10-18.

Bamman, David, and Gregory Crane. “Computational Linguistics and Classical Lexicography.” *Digital Humanities Quarterly*, 3 (January 2009). http://www.digitalhumanities.org/dhq/vol/3/1/000033.html

Clough, Paul D., Neil Ireson, and Jennifer Marlow. "Extending Domain-Specific Resources to Enable Semantic Access to Cultural Heritage Data." *Journal of Digital Information* 10 (2009).

Crane, Gregory. “Generating and Parsing Classical Greek.” *Literary & Linguistic Computing*, 6 (January 1991): 243-245.

Grover, Claire, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. "Use of the Edinburgh Geoparser for Georeferencing Digitized Historical Collections." *Physical and Engineering Sciences* 368 (August 2010): 3875-3889. http://journals.tdl.org/jodi/article/view/698

Packard, David W. “Computer Assisted Morphological Instruction of Ancient Greek.” *Computational and Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics*, 2, (1973): 343-356.

Waltinger, Ulli and Alexander Mehler. "Who Is It? Context Sensitive Named Entity and Instance Recognition by Means of Wikipedia." *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, December 2008, 381-38.